A Brief Survey on Different Privacy Preserving Techniques

Aparna Shinde
Technocrats
Institute Of Technology
RGPV University
Department of
Information Technology
Bhopal, Madhya
Pradesh, India

Khushboo Saxena
Asst. Professor
Technocrats
Institute Of Technology
RGPV University
Department of
Information Technology
Bhopal, Madhya
Pradesh, India

Amit Mishra
Technocrats
Institute Of Technology
RGPV University
Department of
Information Technology
Bhopal, Madhya
Pradesh, India

Shiv K. sahu, PhD
Assoc. Prof. & H.O.D
Technocrats
Institute Of Technology
RGPV University
Department of
Information Technology
Bhopal, Madhya
Pradesh, India

ABSTRACT

As data mining is used to extract valuable information from large amount of data. But this is harmful in some cases so some kind of protection is required for sensitive information. So privacy preserving mining is emerge with the goal to provide protection from mining. There are many research branches in this area. This paper focus on analyzing different techniques of privacy persevering and specify there requirement for special type of cases.

Keywords

Privacy Preserving Mining, Data Perturbation, Aggregation, Data Swapping.

1. INTRODUCTION

As the data miners are gathering information from the large dataset base on useful patterns, trends, etc. This is useful for helping crime understanding, any kind of terrorist activity can also be learn by the data mining approach. With this bright side of the data mining if miners opt to find information about individual then it lead to harm the privacy of the person, place, class. So it is required to provide privacy from such kind of miners activity by applying privacy preserving mining techniques.

This is very useful for the security of data which contain information about the individual, financial information of family or any class. So make some changes on the dataset that is modify or rearrange present information in the dataset so that miner not reach to concern person. Hence many approaches of privacy preserving mining are used for

preserving the information at various level [3, 4]. Information of the individual can be find by observing repeated pattern present in the dataset, then perturb that data by using different methods such as association rules, suppression, swapping, etc.

Miner can access the information, when data is place on the server , so many researchers are working for the access of the data. If data is successfully access then it is possible for miner to get all present. Considering this problem people are working to give more security against large number of privacy attacks. So before sending the data on the public server, data get perturb so that negative data is suppress and it will not affect the overall privacy [5]. So it is required that data protection is done in prior steps by hiding important information such as name, address, date of birth, mobile number of person, etc. But this kind of protection is not sufficient as it directly or indirectly fetch information by using data mining algorithm from the raw data. In order to understand thought from the literature data mining can be apply and if those thought lead to unfair activity then privacy preserving mining is used to preserve those information.

2. PRIVACY PROTOCOL

Mainly three protocols are used for building a privacy-preserving data mining system. The three protocols entities are shown below[2].

2.1 Data Collection Protocol

Data collection protocol manages privacy during data transmission between the data providers to the data warehouse server. It provide the minimum private information to build accurate data mining models and provide only that part of the information to the data warehouse server.

Basic requirements for the data collection protocol; First, it must be scalable; because a data warehouse server can deal with thousands of data providers like online survey system. Second, data provide should be provided data mining at lower cost to increase their participation. Lastly, the protocol must be robust; it must produce relatively accurate data mining results while protecting data provider's privacy, even if data providers have lacking consistency. For example, if data which is provided by an online survey system deviate from the protocol or submit meaningless data, then it must be control the influence of such erroneous behavior and ensure that global data mining results remain sufficiently accurate. Figure-1 shows data collection protocol taxonomy based on two data collection methods.

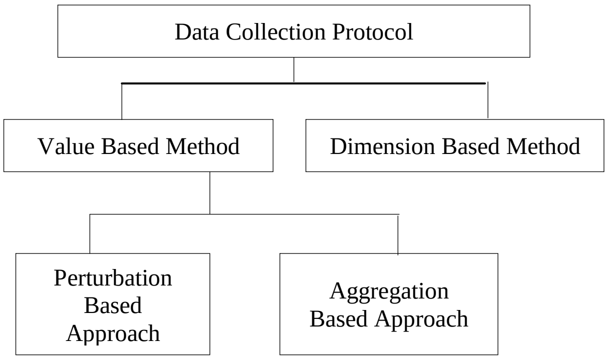

Fig 1: Data Collection Protocol Taxonomy

2.1.1 Value Based Method

With the value-based method, a data provider manipulates the value of each data attribute or item independently using either *perturbation-based* approach or *aggregation-based* approach. The *perturbation-based* approach, noise is directly added to the original data values, such as changing age 25 to 35 or Texas to London. The *aggregation-based* approach, generalization of data is done according to the relevant domain hierarchy, such as changing age 27 to age range 25-30 or Texas to the UK.

The *perturbation-based a*pproach is recommended for random data, while the *aggregation-based* approach depend on knowledge of the domain hierarchy[2], but can be effective in guaranteeing the data's anonymity *k*-anonymity, means that each perturbed data record is not able to be identified from the perturbed values of at least *k*-1 other data record.

The value-based method assumes that it would be difficult, but not impossible, the original data from manipulated values can be discovered from data warehouse servers but that the server would still be able to return the original data distribution from the perturbed data. So easily construct the accurate data mining models.

2.1.2 Dimension Based Method

With the dimension based method data to be mined usually has many attributes or dimension. It removes the private information from the original data by reducing the numbers of dimensions. Construction of data mining servers based on dimension based method results in information loss.

2.2 Inference Control

Inference control protects privacy between the data warehouse server and data mining servers.

2.3 Information sharing

Information sharing gives the control on information which is shared by the data mining servers in different systems.

An aim of these protocols is to produce minimum private information with accuracy for data mining from one entity to another to build accurate data mining models.

3. PRIVACY PRESERVING TECHNIQES

Public concern is mainly caused by the so-called secondary use of personal information without the consent of the subject. In other words, users feel strongly that their personal information should not be sold to other organizations without their prior consent. The majority of respondents in society are concerned about the possible misuse of their personal information. Also shows that, when it comes to the confidence that their personal information is properly handled, consumers have most trust in banks and health care providers and the least trust in credit card agencies and internet companies.

Classification of Privacy preserving techniques can be based on the protection methods used by them which is shown in Figure-2.

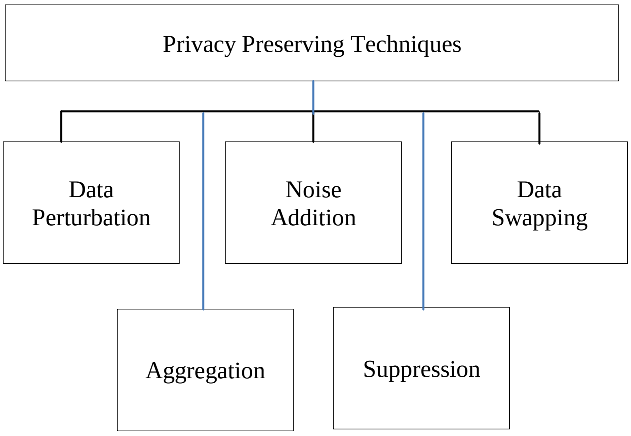

Fig 2: Different Privacy preserving Techniques

3.1 Data Perturbation

Data is directly modified in this technique so it come under data modification category. It is a category of data modification approaches which protect the sensitive data from intruders. Here selected portion of the dataset is consider as the sensitive information which need to be hide by modifying those values or information. So released data is contain inaccurate data where sensitive information is modify. While doing modification it is required to do perturbation in the information having same statistics, as different values get directly act as outliers. So perturbation divide into two main category first is probability distribution approach and the other is value distortion approach. The approach of probability distribution, replaces the data with same data from the distribution of value present in original.

3.2 Noise Addition

This technique is apply on numeric data where some noise producing function can be used to produce noise such as Gaussian function. Here data quality is maintain by the technique so it look like original, while privacy of the information is maintain. [8]. The underlying distributions of a perturbed data set can be unpredictable if the distributions of the corresponding original data set and/or the distributions of the added noise is not multivariate normal. In such a case responses to queries involving percentiles, sums, conditional means etc. Some noise addition techniques, Probabilistic Perturbation Technique, Random Perturbation Technique, All Leaves Probabilistic Perturbation Technique.

3.3 Data Swapping

Data swapping techniques mainly appeal of the method was it keeps all original values in the data set, at the same time the record re-identification is very difficult [1]. Data swapping means replaces the original data set by another one. Here some original values belonging to a sensitive attribute are exchanged between them. This swapping can be done in a way so that the t-order statistics of the original data set are preserved. A t-order statistic is a statistic that can be generated from exactly t attributes. A new concept called approximate data swap was introduced for practical data swapping. It computes the t-order frequency table from the original data set, and finds a new data set with approximately the same t-order frequency. The elements of the new data set are generated one at a time from a probability distribution constructed through the frequency table. The frequency of already created elements and a possible new element is used in the construction of the probability distribution. Inspired by existing data swapping techniques used for statistical databases a new data swapping technique has been introduced for privacy preserving data mining, where the requirement of preserving t-order statistics has been relaxed. The technique emphasizes the pattern preservation instead of obtaining unbiased statistical parameters [2]. It preserves the most classification rules and also obtained different classification algorithms. As the class is typically a categorical attribute containing just two different values, the noise is added by changing the class in a small number of records. It can be achieved by randomly shuffling the class attributes values belonging to heterogeneous leaves of a decision tree.

3.4 Aggregation

Generalization of the available data is also term as aggregation. As this provide privacy of individual information before the release by replacing a group of information with a single. In other words aggregation replace k number of session by its representative session. Such as attribute value in the dataset is derived by taking the average of the bunch of same attribute information. Now this raise one new issue where replacement of k number of original records by a aggregated one make information loss. So in order to reduce this loss of information clustering of the records is required where size of cluster get reduce for decreasing the information loss [3]. But by doing this intruder can estimate of the balance information loss in the data. So overall risk of the of the disclosed data is remain same. Other method of aggregation or generalization is transformation of attribute values. For example, an exact date of birth can be replaced by the year of birth; an exact salary can be replaced rounded in thousands. Such a generalization makes an attribute values less informative. Therefore, a use of excessive extent of generalization can make the released data useless. For example, if an exact date of birth is replaced by the century of birth then the released data can become useless to data miners.

3.5 Suppression

In suppression technique, sensitive data values are deleted or suppressed prior to the release of a data [4]. This technique is used to protect an individual privacy from intruder's attempts to accurately predict a suppressed value. A Sensitive value is predicted by an intruder through various approaches. For example, a built classifier on a released data set can be used in an attempt to predict a suppressed attribute value. Therefore sufficient number of attribute values should be suppressed in order to protect privacy. However, suppression of attribute values results in information loss. An important issue in suppression is to minimize the information loss by minimizing the number of values suppressed. For some applications like a medical diagnosis the suppression is preferred over noise addition in order to reduce the chance of having misleading patterns in the perturbed data set.

By specifying the minimum confidence and support specific items from the dataset can be hide. This can be done by removing or replacing the items from the set then check the minimum support and confidence of that item. In this way by association rule one can implement privacy preserving. Table1 shows comparison of all these privacy preserving techniques.

Table 1: Comparison of Privacy Preserving Techniques

Technique	Advantage	Disadvantage
Data Perturbation	Perturb both Numeric and text data	complexity is high
Noise Addition	Low complexity	Only Numeric value is perturb
Swapping	No data loss	Pattern can be mined easily
Aggregation	Irrepressible for intruders	Not applicable for all kind of data
Suppression	Zero privacy risk	Heavy information loss

4. CONCLUSION & FUTURE SCOPE

Mining information from the data is the primary requirement of the data mining out of which privacy preserving mining is opening new emerging field which preserve knowledge from the data. Paper detailed various methods like perturbing, swapping, etc. for privacy preserving, where each has its own importance. In future data perturbation technique is used to perturb data as compare to other techniques. In this paper comparison of different privacy preserving techniques are done.

5. REFERNCES

[1] FoscaGiannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases*" In IEEESystems Journal, VOL. 7, NO. 3, SEPTEMBER 2013*, pp. 385-395.

[2] N. Zhang and W. Zhao, "Privacy Preserving Data Mining Systems " In IEEE Computer society, 2007 pp. 52-58.

[3] W.K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "Security in outsourcing of association rule mining," in *Proc. Int. Conf.*Very Large Data Bases, 2007, pp. 111–122.

[4] K.Sathiyapriya and Dr. G.SudhaSadasivam, " A Survey on Privacy Preserving Association Rule Mining", In IJKDP Vol.3 No 2– March-2013, pp 119-131.

[5] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc.ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 439–450.

[6] ManopPhankokkruad, " Association Rules for Data Mining in Item Classification Algorithm : Web Service Approach ", In IEEE, 2012 pp. 463-468.

[7] D.Narmadha, G.NaveenSundar and S.Geetha,"*A Novel Approach to Prune Mined Association Rules in Large Databases‖*", IEEE, 2011 pp.

[8] T zung -Pei, Hong Kuo-Tung Yang, Chun-Wei Lin and Shyue-Liang Wang, "Evolutionary privacy preserving in data mining ",In IEEE World Automation Congress conference , 2010 pp.

[9] Z. Yang and R. N. Wright. “Privacy-preserving computation of bayesian networks on vertically partitioned data.” In IEEE Trans. on Knowledge and Data Engineering , 2006, pp.1253–1264.

[10] Enabling Multilevel Trust in Privacy Preserving Data Mining Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012.

[11] Survey on Privacy Preserving Data Mining Haitao Liu and Jing GeUniversity of Illinois at Urbana-Champaign, UrbanaIL, 61801 USA.

[12] Privacy- Preserving Data Publishing for MultipleNumerical Sensitive Attributes QinghaiLiu, Hong Shen_, and Yingpeng Sang. TSINGHUA SCIENCE AND TECHNOLOGY Volume 20, Number 3, Jun 2015